DATA ARTICLE

REVISED Matched molecular pair-based data sets for computer-aided medicinal chemistry [v2; ref status: indexed, http://f1000r.es/309]

Ye Hu, Antonio de la Vega de León, Bijun Zhang, Jürgen Bajorath

Department of Life Science Informatics,B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Bonn, D-53113, Germany

v2 **First Published:** 04 Feb 2014, **3**:36 (doi: 10.12688/f1000research.3-36.v1)
Latest Published: 21 Feb 2014, **3**:36 (doi: 10.12688/f1000research.3-36.v2)

Abstract

Matched molecular pairs (MMPs) are widely used in medicinal chemistry to study changes in compound properties including biological activity, which are associated with well-defined structural modifications. Herein we describe up-to-date versions of three MMP-based data sets that have originated from in-house research projects. These data sets include activity cliffs, structure-activity relationship (SAR) transfer series, and second generation MMPs based upon retrosynthetic rules. The data sets have in common that they have been derived from compounds included in the ChEMBL database (release 17) for which high-confidence activity data are available. Thus, the activity data associated with MMP-based activity cliffs, SAR transfer series, and retrosynthetic MMPs cover the entire spectrum of current pharmaceutical targets. Our data sets are made freely available to the scientific community.

Article Status Summary

Referee Responses

Referees	1	2	3	4
v1 published 04 Feb 2014	✓ report	✓ report	✓ report	✓ report
v2 published 21 Feb 2014 REVISED			✓ report	

1 **Ajay Jain**, University of California San Francisco USA

2 **Peter Ertl**, Novartis Institutes for Biomedical Research Switzerland

3 **Patrick Walters**, Vertex Pharmaceuticals Inc. USA

4 **Shana Posy**, Bristol-Myers Squibb USA

Latest Comments

Jürgen Bajorath, Department of Life Science Informatics, B-IT and LIMES Institutes, Rheinische Friedrich-Wilhelms-Universität Bonn, Germany
20 Feb 2014 (V1)

Jürgen Bajorath, Department of Life Science Informatics, B-IT and LIMES Institutes, Rheinische Friedrich-Wilhelms-Universität Bonn, Germany
18 Feb 2014 (V1)

Corresponding author: Jürgen Bajorath (bajorath@bit.uni-bonn.de)

How to cite this article: Hu Y, de la Vega de León A, Zhang B *et al.* (2014) Matched molecular pair-based data sets for computer-aided medicinal chemistry [v2; ref status: indexed, http://f1000r.es/309] *F1000Research* 2014, **3**:36 (doi: 10.12688/f1000research.3-36.v2)

Grant information: The author(s) declared that no grants were involved in supporting this work.

Competing Interests: No competing interests were disclosed.

First Published: 04 Feb 2014, **3**:36 (doi: 10.12688/f1000research.3-36.v1)
First Indexed: 17 Feb 2014, **3**:36 (doi: 10.12688/f1000research.3-36.v1)

REVISED **Amendments from Version 1**

The version of the ChEMBL database used in our analysis and more technical information concerning calculations and toolkits have been provided. Standard MMPs and RECAP-MMPs have been compared in more detail and MMP-cliffs with a potency difference of at least one order of magnitude have also been determined. Table 2 has been updated and a new Figure 4 has been added.

The data sets have been updated. In the files of MMP-cliffs and RECAP-MMPs, a column "NumOfCuts" has been added indicating the origin of chemical transformations. Target names have been added in all files. Compound activities have been incorporated in files of RECAP-MMPs. For transfer series, substituted fragments have been provided.

See referee reports

Introduction

The matched molecular pair (MMP) concept is widely applied in medicinal chemistry[1–4]. An MMP is defined as a pair of compounds that are only distinguished by a structural modification at a single site[1], i.e., the exchange of a substructure, termed a chemical transformation[5]. MMPs are attractive tools for computational analysis because they can be algorithmically generated and they make it possible to associate defined structural modifications at the level of compound pairs with chemical property changes, including biological activity[2–4]. MMPs are usually chemically intuitive and easily accessible, which helps to bridge the gap between computational analysis and the practice of medicinal chemistry.

In the context of different studies, we have systematically generated MMPs through the mining of publicly available compound activity data. All possible MMPs have been derived from compounds active against currently available pharmaceutical targets. Then, MMPs have been used to explore structure-activity relationships (SARs) on a large-scale and from different viewpoints.

In a previous data article, we have reported and made publicly available a number of different data sets and computational tools developed in our laboratory[6]. Here we describe three recently developed MMP-based data structures, which should be of interest for SAR analysis and compound design, and we also provide up-to-date versions of the corresponding data sets. It is anticipated that these data sets will be helpful as a resource for computer-aided medicinal chemistry applications. The data sets include MMP-based activity cliffs (i.e., MMP-cliffs), SAR transfer series, and MMPs derived on the basis of retrosynthetic fragmentation rules and were derived from all bioactive compounds currently available in the ChEMBL database (release 17)[7,8]. Only high-confidence activity data (as specified below) were considered. MMP-cliffs, SAR transfer series, and retrosynthetic MMPs provide comprehensive sources of SAR information. In addition, retrosynthetic MMPs are thought to increase the utility of computational MMP analysis for practical chemistry efforts because these second generation MMPs consider reaction information during molecular fragmentation, which sets them apart from standard MMPs originating from systematic fragmentation of all possible exocyclic single bonds in a molecule (as detailed below).

Materials and methods

Concepts

(1) *Activity cliffs* are generally defined as pairs or groups of compounds that are structurally similar and have large differences in potency[9–11]. Accordingly, activity cliffs usually have high SAR information content (because small chemical changes in similar or analogous compounds lead to large potency effects). The assessment of activity cliffs requires clearly defined similarity and potency difference criteria[9–11]. The formation of an MMP can be considered as a similarity criterion, which is similarity metric-free and often chemically more intuitive than the use of calculated molecular similarity[11,12]. MMP formation as a similarity criterion has led to the introduction of MMP-cliffs[12]. For MMP-cliffs, a difference in potency of at least two orders of magnitude between cliff-forming compounds was set as a potency difference criterion[12]. Figure 1 shows exemplary MMP-cliffs.

(2) *SAR transfer* can be rationalized in different ways. For example, a compound series might display similar potency progression against two different targets[13]. Alternatively, two different compound series with corresponding analogs, i.e., series having different core structures and containing compounds with pairwise corresponding substitutions, might display similar potency progression against a given target[14]. Such *SAR transfer series* displaying similar target-specific SAR behavior are often sought after in medicinal chemistry as alternative compounds for optimization. Here we focus on these target-based SAR transfer series. Figure 2 shows an example.

(3) Computational generation of MMPs typically involves molecular fragmentation through the systematic deletion of exocyclic single bonds[5]. Hence, the resulting fragments representing a molecular core and substituent are not derived considering chemical reactions. Accordingly, a transformation relating MMP-forming compounds to each other might not necessarily be interpretable from a synthetic perspective. Hence synthetic accessibility of MMPs might be further improved by considering the reaction information during molecular fragmentation. This has been accomplished by applying the well-known retrosynthetic combinatorial analysis procedure (RECAP) rules[15], leading to the introduction of *RECAP-MMPs*[16]. Representative examples are shown in Figure 3. In addition, exemplary differences between standard MMPs and RECAP-MMPs are illustrated in Figure 4.

MMP generation

For the generation of MMP-cliffs, SAR transfer series, and RECAP-MMPs, transformation size restrictions that limit transformations to meaningful chemical substitutions were introduced[12]. Specifically, the common core structure had to be at least twice the size of each exchanged substructure. Furthermore, the difference in size of the exchanged fragments was limited to at most eight non-hydrogen atoms and the maximal size of an exchanged fragment was set to 13 non-hydrogen atoms[12]. Therefore, the largest permitted transformations included, for example, the addition of a substituted ring to a compound or the replacement of a five- or six-membered ring with a substituted condensed two-ring system (with a maximum of 13 atoms). All possible transformation size-restricted MMPs and

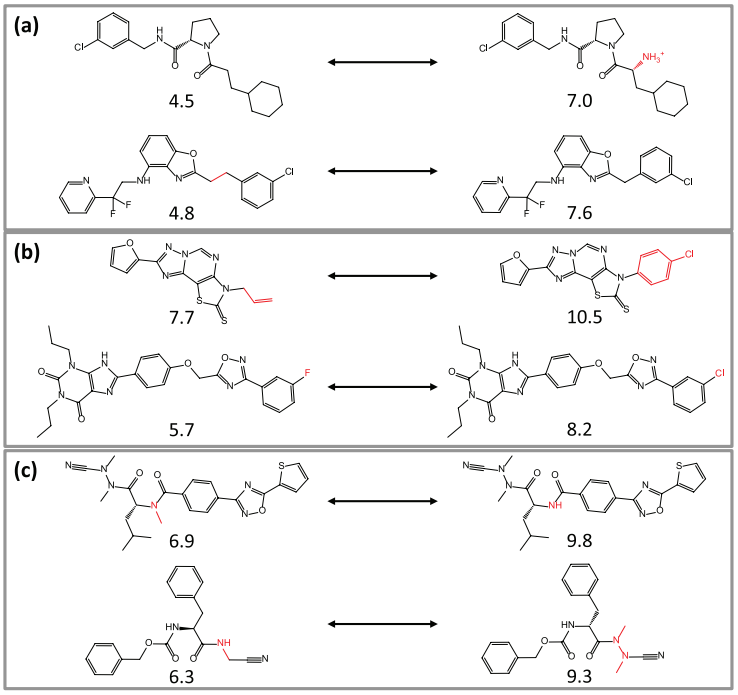

Figure 1. MMP-cliffs. Six representative MMP-cliffs for three targets belonging to different target families are shown; (**a**) muscarinic acetylcholine receptor M3, (**b**) serine/threonine-protein kinase c-TAK1, (**c**) matrix metalloproteinase-2. The pK_i value of each compound is provided and the structural differences between cliff-forming compounds are highlighted in red.

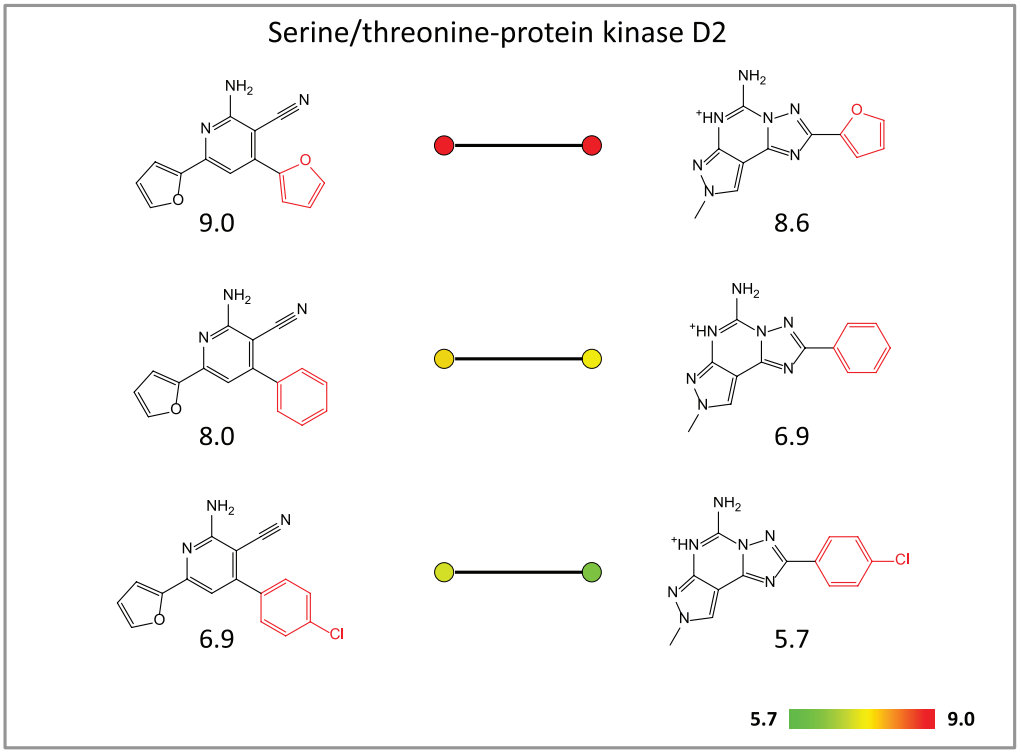

Figure 2. SAR transfer series. An exemplary target-based SAR transfer series is shown. Compound pairs are arranged in the order of increasing potency (from the bottom to the top). Potency progression is monitored by corresponding pairs of color-coded dots using a continuous color spectrum from green (lowest potency value (pK_i = 5.7) in the compound data set), over yellow to red (highest potency value; pK_i = 9.0). The pK_i value of each compound is provided. The core structures are drawn in black and the substituents in red. The compounds are active against serine/threonine-protein kinase D2.

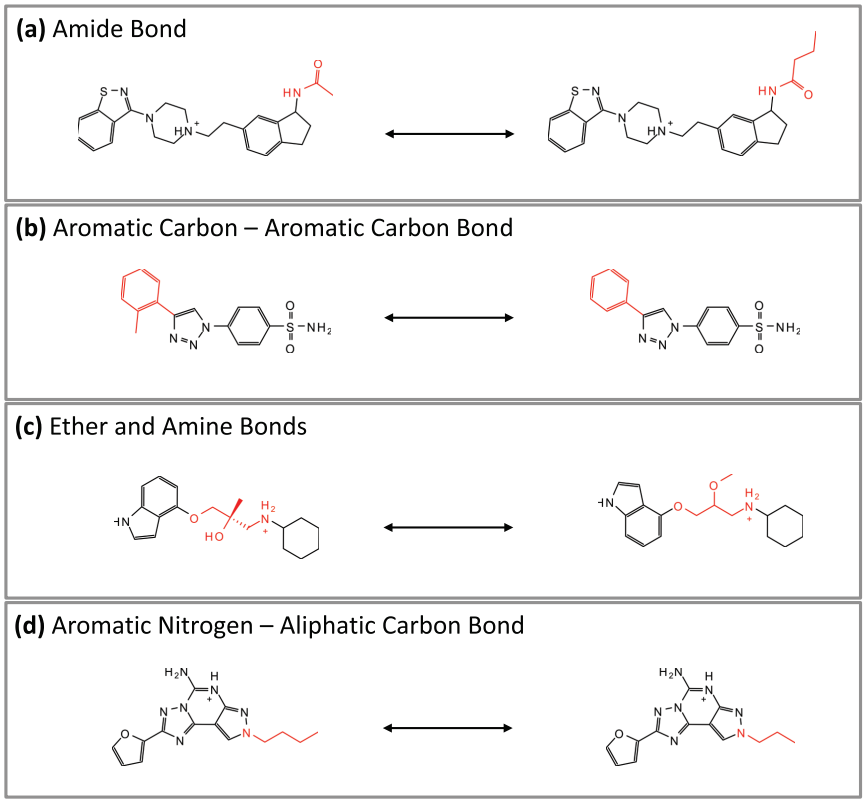

Figure 3. RECAP-MMPs. In (**a**)–(**d**), four exemplary RECAP-MMPs representing different retrosynthetic rules are shown. For each RECAP-MMP, the chemical transformation is highlighted in red.

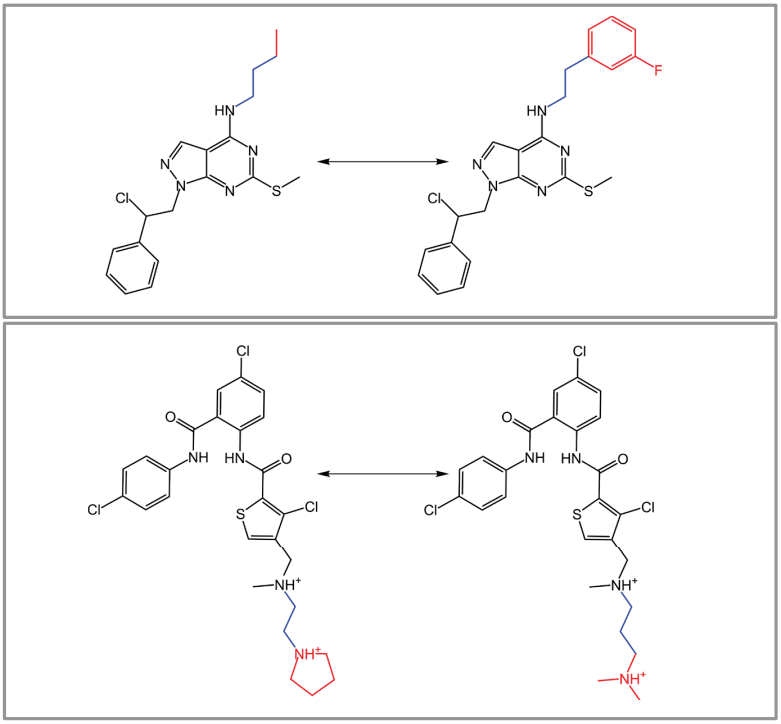

Figure 4. Standard MMPs vs. RECAP-MMPs. Two pairs of compounds that form both standard MMPs and RECAP-MMPs are shown. For each pair, the structural differences between compounds are highlighted. The chemical transformation associated with the standard MMP is colored in red, while the transformation of the RECAP-MMP corresponds to the combination of fragments colored in red and blue.

RECAP-MMPs were calculated using an in-house implementation of the algorithm by Hussain and Rea[5] that utilizes the OpenEye toolkit[17].

Compounds and activity data

Compound data were taken from the latest version of ChEMBL (release 17)[7,8]. Only compounds with direct interactions (i.e., target relationship type "D") against human targets at the highest confidence level (target confidence score 9) were selected. Two types of potency measurements were separately considered, i.e., K_i (equilibrium constant) and IC_{50} (half-maximal inhibition concentration) values. In order to ensure high data confidence, inactive or inconclusive compounds and compounds with approximate measurements such as ">", "<", or "~" were not considered. For compounds with multiple measurements against the same target, the geometric mean was calculated as the final potency annotation, provided that all values fell within one order of magnitude; otherwise, the compound was discarded. All qualifying compounds were further organized into target sets. A total of 661 and 1203 target sets (consisting of compounds with reported specific activity against a given target) were collected for the K_i- and IC_{50}-based subsets, respectively, as reported in Table 1. The target sets contained a total of 45,353 and 95,685 compounds and 77,421 and 135,291 potency measurements for the K_i and IC_{50} subsets, respectively. These target sets provided the basis for the generation of all MMPs.

Table 1. Data sets.

Number of	K_i	IC_{50}
Targets	661	1203
Compounds	45,353	95,685
Measurements	77,421	135,291

For the K_i and IC_{50} subsets from the latest version of ChEMBL (release 17), the total numbers of targets, compounds, and corresponding potency measurements are reported.

Results

As a follow-up on the original publications in which MMP-cliffs[12], SAR transfer series[14], and RECAP-MMPs[16] were introduced, all corresponding data sets have been re-generated on the basis of ChEMBL release 17, hence providing up-to-date versions for release. Separate data subsets have been generated for different types of well-defined potency measurements (i.e., assay-dependent IC_{50} vs. assay-independent K_i values) to avoid inconsistencies due to simultaneous consideration of different potency measurements that cannot be directly compared.

MMP-cliffs

Figure 1 illustrates small chemical changes in compound pairs leading to large potency differences that are captured by MMP-cliffs. For ease of structural interpretation, we currently prefer MMP-based activity cliff representations compared to alternative representations that rely on calculated similarity values[11]. Table 2 provides the MMP-cliff statistics for the current data set. On the basis of K_i and IC_{50} measurements, more than 20,000 and 25,000 MMP-cliffs were obtained, respectively, requiring an at least 100-fold difference in potency between cliff-forming compounds. The MMP-cliffs corresponded to ~5% of all MMPs that were generated from ChEMBL compounds with high-confidence activity data. They covered 293 and 500 different targets on the basis of K_i and IC_{50} measurements, respectively. In addition to the more conservative potency difference cutoff, MMP-cliffs were also identified when a less stringent criterion was applied, i.e., two compounds forming an MMP were required to have a potency difference of at least one order of magnitude. In this case, as reported in Table 2, nearly 99,000 and more than 126,000 MMP-cliffs were detected in

Table 2. MMP and MMP-cliff statistics.

Number of		K_i	IC_{50}
MMPs		385,777	537,848
Targets with MMPs		467	929
MMP compounds		40,454 (89.2%)	80,744 (84.4%)
ΔPotency ≥ 1 OoM	**MMP-cliffs**	98,608	126,464
	% MMP-cliffs	25.6%	23.5%
	Targets with MMP-cliffs	392	726
	MMP-cliff compounds	29,976 (66.1%)	50,413 (52.7%)
ΔPotency ≥ 2 OoM	**MMP-cliffs**	20,073	25,297
	% MMP-cliffs	5.2%	4.7%
	Targets with MMP-cliffs	293	500
	MMP-cliff compounds	11,760 (25.9%)	16,816 (17.6%)

For the K_i- and IC_{50}-based compound subsets, the number of MMPs, the number of targets for which MMPs were obtained, and the number (and ratio) of compounds that formed MMPs are reported. In addition, the number and proportion of MMP-cliffs derived from all MMPs with potency difference (ΔPotency) of at least one order (1 OoM) or two orders of magnitude (2 OoM) are provided, respectively, as well as the number of targets for which MMP-cliffs were obtained and the number (and ratio) of cliff-forming compounds.

392 and 726 targets for the K_i and IC_{50} subsets, respectively. The proportion of MMP-cliffs increased to approx. 25%.

SAR transfer series

SAR transfer series are best rationalised as pairs of compound series active against the same target that have distinct core structures, and consist of corresponding pairs of analogs, as illustrated in Figure 2 for a small series with three pairs. Different from the original analysis of target-based SAR transfer[14] that was based upon MMPs without transformation size restrictions, the current analysis has been carried out on the basis of size-restricted MMPs. This modification further supports SAR exploration (because only small chemical changes are considered) and explains a reduction in series numbers compared to the original publication. In Table 3, the numbers of different series available for the current data set are reported. A total of 1270 and 2109 matching series were obtained from the K_i and IC_{50} subsets, respectively. Matching series met the structural requirement of consisting of at least three pairs of corresponding analogs. In addition, the potency values of compounds associated with individual series had to span at least two orders of magnitude. From these pre-selected matching series, 157 (K_i) and 513 (IC_{50}) SAR transfer series with at least approximate potency progression and activity against 42 and 54 targets, respectively, were obtained. A subset of 60 (K_i) and 322 (IC_{50}) SAR transfer series displayed strictly corresponding (regular) potency progression (often over different potency ranges)[14]. These series were active against 23 (K_i) and 27 (IC_{50}) different targets. The size of SAR transfer series with approximate and regular potency progression ranged from three to 12 corresponding pairs of analogs. On average, the SAR transfer series consisted of three to four pairs.

Table 3. Target-based SAR transfer series statistics.

Number of	K_i	IC_{50}
Matching series	1270	2109
T_SAR-TS	157	513
Targets with T_SAR-TS	42	54
T_SAR-TS-RP	60	322
Targets with T_SAR-TS-RP	23	27

For the K_i and IC_{50} subsets, the number of qualifying matching compound series is reported. In addition, the number of target-based SAR transfer series with at least approximate potency progression (T_SAR-TS), the subset of SAR transfer series with regular potency progression (T_SAR-TS-RP), and the corresponding numbers of targets are given.

RECAP-MMPs

The replacement of systematic fragmentation of exocyclic single bonds with a set of 13 retrosynthetic rules for MMP generation reduced the number of MMPs that were obtained by more than half. RECAP-MMP numbers are reported in Table 4. However, (perhaps surprisingly) large numbers of RECAP-MMPs remained for further consideration and assessment of synthetic feasibility. From the K_i and IC_{50} subsets, nearly 170,000 and more than 240,000 RECAP-MMPs were obtained with activity against 371 and 778 targets, respectively. Examples are shown in Figure 3.

Table 4. RECAP-MMP statistics.

Number of	K_i	IC_{50}
RECAP-MMPs	169,889	240,322
Targets with RECAP-MMPs	371	778
RECAP-MMP compounds	28,529 (62.9%)	53,917 (56.3%)

For the K_i and IC_{50} subsets, the number of RECAP-MMPs, the number of targets for which RECAP-MMPs were obtained, and the number (and ratio) of compounds that formed RECAP-MMPs are reported.

Data availability

All MMP-cliffs, SAR transfer series, and RECAP-MMPs are provided in canonical SMILES representation[18] on a per-target basis separately for the K_i and IC_{50} subsets. The canonical SMILES representation of compounds was calculated using the Molecular Operating Environment[19] on the basis of standardized molecular structures by removing solvents or ions and rebalancing protonation states. Furthermore, the canonical SMILES representation of key fragments (cores) and chemical transformations derived from MMPs and RECAP-MMPs was generated using the OpenEye toolkit[17].

ZENODO: Detailed data sets of MMP-cliffs, SAR transfer series, RECAP-MMPs and compound activities, doi: 10.5281/zenodo.8418[20].

Summary

We have described new and up-to-date MMP-based data sets comprising activity cliffs, SAR transfer series, and second generation retrosynthetic MMPs that have been systematically generated from currently available public domain compounds with high-confidence activity data. Hence, these data sets are comprehensive and have broad target coverage. They are made available without restrictions to the scientific community to aid in SAR analysis, compound design, and other medicinal chemistry applications. It is hoped that these data sets might be of interest and useful to many investigators in this field and catalyse further research efforts.

Author contributions

JB designed the study, YH, AVL, and BZ collected and organized the data, JB and YH wrote the manuscript, all authors examined the manuscript and agreed to the final content.

Competing interests

No competing interests were disclosed.

Grant information

The author(s) declared that no grants were involved in supporting this work.

Acknowledgements

We thank OpenEye Scientific Software, Inc., for the free academic license of the OpenEye Toolkits.

References

1. Kenny PW, Sadowski J: **Structure modification in chemical databases.** In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2005; 271–285.
Publisher Full Text
2. Griffen E, Leach AG, Robb GR, *et al.*: **Matched molecular pairs as a medicinal chemistry tool.** *J Med Chem.* 2011; **54**(22): 7739–7750.
PubMed Abstract | Publisher Full Text
3. Wassermann AM, Dimova D, Iyer P, *et al.*: **Advances in computational medicinal chemistry: matched molecular pair analysis.** *Drug Dev Res.* 2012; **73**(8): 518–527.
Publisher Full Text
4. Dossetter AG, Griffen EJ, Leach AG: **Matched molecular pair analysis in drug discovery.** *Drug Discov Today.* 2013; **18**(15–16): 724–731.
PubMed Abstract | Publisher Full Text
5. Hussain J, Rea C: **Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets.** *J Chem Inf Model.* 2010; **50**(3): 339–348.
PubMed Abstract | Publisher Full Text
6. Hu Y, Bajorath J: **Freely available compound data sets and software tools for chemoinformatics and computational medicinal chemistry applications [v1; ref status: indexed, http://f1000r.es/Mu9krs].** *F1000Res.* 2012; **1**: 11.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Gaulton A, Bellis LJ, Bento AP, *et al.*: **ChEMBL: a large-scale bioactivity database for drug discovery.** *Nucleic Acids Res.* 2012; **40**(Database issue): D1100–D1107.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Bento AP, Gaulton A, Hersey A, *et al.*: **The ChEMBL bioactivity database: an update.** *Nucleic Acids Res.* 2014; **42**(1): D1083–D1090.
PubMed Abstract | Publisher Full Text
9. Stumpfe D, Bajorath J: **Exploring activity cliffs in medicinal chemistry.** *J Med Chem.* 2012; **55**(7): 2932–2942.
PubMed Abstract | Publisher Full Text
10. Stumpfe D, Hu Y, Dimova D, *et al.*: **Recent progress in understanding activity cliffs and their utility in medicinal chemistry.** *J Med Chem.* 2014; **57**(1): 18–28.
PubMed Abstract | Publisher Full Text
11. Hu Y, Stumpfe D, Bajorath J: **Advancing the activity cliff concept [v1; ref status: indexed, http://f1000r.es/1wf].** *F1000Res.* 2013; **2**: 199.
Publisher Full Text | Free Full Text
12. Hu X, Hu Y, Vogt M, *et al.*: **MMP-Cliffs: systematic identification of activity cliffs on the basis of matched molecular pairs.** *J Chem Inf Model.* 2012; **52**(5): 1138–1145.
PubMed Abstract | Publisher Full Text
13. Zhang B, Hu Y, Bajorath J: **SAR transfer across different targets.** *J Chem Inf Model.* 2013; **53**(7): 1589–1594.
PubMed Abstract | Publisher Full Text
14. Zhang B, Wassermann AM, Vogt M, *et al.*: **Systematic assessment of compound series with SAR transfer potential.** *J Chem Inf Model.* 2012; **52**(12): 3138–3143.
PubMed Abstract | Publisher Full Text
15. Lewell XQ, Judd DB, Watson SP, *et al.*: **RECAP--retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry.** *J Chem Inf Comput Sci.* 1998; **38**(3): 511–522.
PubMed Abstract | Publisher Full Text
16. de la Vega de León A, Bajorath J: **Matched molecular pairs derived by retrosynthetic fragmentation.** *Med Chem Commun.* 2014; **5**(1): 64–67.
Publisher Full Text
17. **OEChem, version 1.7.7**, OpenEye Scientific Software, Inc., Santa Fe, NM, USA. 2012.
Reference Source
18. Weininger D: **SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules.** *J Chem Inf Comput Sci.* 1988; **28**(1): 31–36.
Publisher Full Text
19. **Molecular Operating Environment (MOE), 2011.10;** Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite#910, Montreal, QC, Canada, H3A 2R7, 2011.
Reference Source
20. Hu Y, de la Vega de León A, Zhang B, *et al.*: **Detailed data sets of MMP-cliffs, SAR transfer series, RECAP-MMPs and compound activities.** 2014.
Data Source

Current Referee Status:

Referee Responses for Version 2

Patrick Walters
Vertex Pharmaceuticals Inc., Cambridge, MA, USA

Approved: 24 February 2014

Referee Report: 24 February 2014
This revised version looks good.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Responses for Version 1

Shana Posy
Research and Development, Bristol-Myers Squibb, Princeton, NJ, USA

Approved: 20 February 2014

Referee Report: 20 February 2014
Hu *et al* have compiled a useful set of matched pair datasets based on the CHEMBL database of biological activity. They describe in a straightforward manner the derivation of the datasets and basic concepts relevant for matched pairs. The following suggested modifications to the data provided would enhance their usefulness and completeness:

1. As a result of switching the method for generating MMPs from cleavage of single bonds to a RECAP-based method, the pool of MMPs now includes substitutions of internal fragments (e.g. in Figure 3c) as well as substitution of a terminal R-group (as in Figure 3 examples a, b, and d). Although both types of MMPs involve replacement of a single structural fragment, it may be desirable for many applications to distinguish between core scaffold replacement and R-group variation. It would therefore be helpful to annotate the datasets to easily separate these two classes of MMPs.

2. Since the authors filter out IC50s/Kis of indeterminate values, it is unclear how compounds that were clearly inactive were processed. Were compounds with IC50s/Kis that could not be quantified due to a flat dose response curve included in the datasets?

3. The authors present a filtered dataset where a number of factors have contributed to rejection of potential MMPs, namely: the difference in size of the exchanged fragments was limited to 8 heavy atoms; the ratio of the common core fragment to the size of each exchanged fragment had to

be>2; and the exchanged fragment could have maximum 13 heavy atoms. While these are reasonable filters to obtain MMPs that truly represent small structural changes, the cutoffs selected are arbitrary and for some targets may exclude MMPs that another user might consider relevant. Rather than providing the final filtered dataset, it would be helpful if the authors would provide the full original datasets with the values of the features used for filtering annotated as extra columns. This would allow maximal flexibility in designing custom MMP sets.

4. In the files that list the RECAP MMPs, key fields are missing that would require the user to retrieve the relevant data from CHEMBL in order to perform any analysis: (a) the Target name (only the target CHEMBL ID is provided); and (b) more importantly, the compound activities are not included.

5. In the files that list the transfer series, for each matched pair the authors provide the two series cores and full compound smiles, but not the substituted fragments.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Patrick Walters
Vertex Pharmaceuticals Inc., Cambridge, MA, USA

Approved: 18 February 2014

Referee Report: 18 February 2014
This paper provides a review of the Bajorath group's recent work on matched molecular pairs (MMP), a technique for exploring structure activity relationships, and identifying chemical transformations that can readily modulate biological activity. The authors focus on recent applications of MMP to large datasets from the publicly accessible ChEMBL database. The paper provides an excellent introduction to those unfamiliar with the MMP technique and with concepts such as activity cliffs. In addition to providing an overview of the recent literature, the authors also provide links to publicly available software and datasets that will provide tutorial materials for those interested in learning more about these powerful techniques. Datasets and software like those described in this paper are valuable resources. A logical next step from this work would be to create interactive tutorials using tools like the iPython Notebook Viewer or knitr.

The presentation is clear, but a few changes may help readers unfamiliar with some of the concepts.

- On p2 the authors refer to "*second generation MMPs*". It would be useful to add a sentence explaining the differences between first and second generation MMPs.

- MMP Cliffs which differ in activity by 10-fold may also be interesting. It would be informative to see the number of examples available with a 1 log vs 2 log difference.

- In the section (3) on page 3, it would be interesting to provide a specific example of how the results of RECAP generated MMPS differ from those generated using a more "traditional" approach.

This paper provides an excellent gateway to a topic that is becoming increasingly more important in drug discovery. The paper should be of interest to computational and medicinal chemists as well as biologists.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Peter Ertl
Cheminformatics, Novartis Institutes for Biomedical Research, Basel, Switzerland

Approved: 17 February 2014

Referee Report: 17 February 2014
The matched molecule pairs approach provides a “chemistry friendly” and intuitive way of expressing relationships among molecules and therefore this manuscript is of importance to all cheminformatics scientists interested in the study of activity cliffs, SAR analysis and in the design of bioactive molecules in general.

The authors extracted several datasets of matched molecule pairs from public databases that could be used as benchmarks in further analysis of this phenomenon. Care has been taken to assure a high quality of data.

The manuscript is well written and the procedure and all results are sufficiently documented and, in addition, all datasets are available for download; therefore I am suggesting only a few minor modifications to the text:

1. Introduction: replace "*the latest release of the ChEMBL*" with the version number.

2. Provide a bit more technical information about the in-house implementation of a molecule fragmentation procedure used to generated matched pairs. Was the procedure implemented entirely in-house, or is it based on a publicly available cheminformatics toolkit? (If this is the case, please cite the respective toolkit).

3. The authors mention that structures in the download file are available as canonical SMILES. The form of canonical SMILES however will depend on the particular program used to generate it. Please specify whether the original ChEMBL SMILES is included or the canonical SMILES was created by another toolkit

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Ajay Jain
Departments of Biopharmaceutical Sciences and Laboratory Medicine, University of California San Francisco, San Francisco, CA, USA

Approved: 10 February 2014

Referee Report: 10 February 2014
The data set described by Hu *et al.* is a large set of carefully curated small molecule matched-molecular pairs (MMPs) with high-quality activity data derived from ChEMBL. The set includes examples of structure-activity cliffs, as well as matched SAR-transfer series, both of which are important in the development and validation of activity prediction algorithms. The availability of the MMP data set will be very valuable to researchers that are focused on methods development. The data should also be of interest to those interested in fundamental questions about molecular activity (e.g. questions about the independence and additivity of activity changes that are linked with substituent changes).

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Article Comments

Comments for Version 1

Author Response

Jürgen Bajorath, Department of Life Science Informatics, B-IT and LIMES Institutes, Rheinische Friedrich-Wilhelms-Universität Bonn, Germany
Posted: 20 Feb 2014

After completing our revision, an additional review has been obtained which we address as follows:

1. In the files of MMP-cliffs and RECAP-MMPs, a column "NumOfCuts" has been added indicating the origin of chemical transformations.

   Single cut indicates that the chemical modification maps to the termini of a molecule, whereas double and triple cuts indicate that the structural changes are at internal parts. It should be noted that changes at termini do not necessarily mean R-group variation (e.g., Figure 3b) and that changes of internal parts do not necessarily mean core scaffold replacement (e.g., Figure 3c).

2. Inactive and inconclusive compounds were not included in our data sets, as stated in the text.

3. Transformation size restrictions were not defined arbitrarily; they are rationalized in the original publication of MMP-cliffs cited in the paper.

4. Target names have been added in all files. Compound activities have been incorporated in files of RECAP-MMPs.

5. For transfer series, substituted fragments have been provided.

In response to the reviewer comments, the data sets have been updated.

We thank the reviewer for the comments ***Competing Interests:*** No competing interests were disclosed.

Author Response

Jürgen Bajorath, Department of Life Science Informatics, B-IT and LIMES Institutes, Rheinische Friedrich-Wilhelms-Universität Bonn, Germany
Posted: 18 Feb 2014

The article has been revised as follows in response to the reviewer comments:

The ChEMBL version number and more technical information concerning toolkits and SMILES calculations have been added. In addition, standard MMPs and RECAP-MMPs have been compared in more detail and differences have been illustrated. Furthermore, MMP-cliffs with a potency difference of at least one order of magnitude have also been determined and reported. Table 2 has been updated and a new Figure 4 has been added.

We thank all three referees for their comments.

Competing Interests: None